DOCUMENT RESUME

ED 338 690 TM 017 515

AUTHOR Tucker, Mary; Taylor, Dianne
TITLE Applying Procrustean Rotation To Evaluate the Generalizability of Research Results.
PUB DATE Jan 91
NOTE 18p.; Paper presented at the Annual Meeting of the Southwest Educational Research Association (San Antonio, TX, January 24-26, 1991).
PUB TYPE Reports - Evaluative/Feasibility (142) -- Speeches/Conference Papers (150)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS *Discriminant Analysis; *Generalizability Theory; Goodness of Fit; Heuristics; Predictor Variables; *Research Methodology; *Statistical Significance; Validity
IDENTIFIERS Empirical Research; *Invariance; *Procrustes Rotation; Research Replication

ABSTRACT

The use of Procrustean rotation as a procedure for assessing the invariance of study results is proposed. Researchers have long relied on significance testing as a measure of judging the worthiness of empirical findings. However, significance testing has come under fire because it does not provide information about the importance or replicability of results. A major misconception is confusing statistical significance testing with reproducibility. Assessing the invariance of study results is a useful alternative. An example is given of an invariance technique following a discriminant analysis. The analysis was calculated from a hypothetical data set with 64 cases and two predictor variables. The rotation technique can be used as a cross-validation procedure, splitting the data from a single sample and comparing the factor vectors from each half. A Procrustean rotation forces orthogonal (uncorrelated) functions of factors to a "best fit" position after setting the factor vectors to unit length to equalize the contribution of each factor vector to the determination of the amount of rotation necessary. The RELATE computer program of D. J. Veldman was used for the necessary calculations. Four tables present data from the analysis. A 20-item list of references is included. (SLD)

APPLYING PROCRUSTEAN ROTATION TO EVALUATE THE GENERALIZABILITY OF RESEARCH RESULTS

Mary Tucker

Nicholls State University

Thibodaux, LA 70310

Dianne Taylor

University of New Orleans

New Orleans, LA 70148

Paper presented at the annual meeting of the Southwest Educational Research Association, San Antonio, TX, January 25, 1991.

ABSTRACT

Researchers have long relied on significance testing as a measure of judging the worthiness of empirical findings. However, in the last two decades, significance testing has come under fire from prominent researchers. Statistical significance testing does not provide information about the importance or the replicability of results. A major misconception is the confusing of statistical significance testing with reproducibility. Thoughtful researchers have begun to place importance on replicability of results. Perhaps one reason for the difficulty in "exorcising the null hypothesis" is that researchers do not feel a suitable substitute has been offered. The present paper offers one such alternative--assessing the invariance of study results. The use of Procrustean rotation as an invariance procedure is the focus of this paper. A concrete example is provided.

Significance testing has long been the measure for judging the worthiness of empirical findings. Adherence to this measure is based on the rationale that "the larger two random samples are, the closer should be their means on any measure of interest, provided the samples are from the same population" (Fish, 1986, pp. 1-2). As Fish (1986) notes, "the logic of statistical significance testing is at first compelling, for it is based on [a] perfectly reasonable assumption" (p. 1). However, in the last two decades, significance testing has come under fire from prominent researchers such as Cronbach (1975), who asserts that "the time has arrived to exorcise the null hypothesis" (p. 124), and Shulman (1970), who maintains that "the time has arrived for educational researchers to divest themselves of the yoke of statistical hypothesis testing" (p. 389).

One reason for the growing disenchantment with statistical significance testing in some quarters is the strong effect that sample size has on the results of a statistical test of the null hypothesis. An example by Thompson (1989) clearly makes this point. Thompson establishes a fixed effect size of 33.6%, considered large in social science research. By using a sample size of 13 cases Thompson is able to create non-significant results, but by increasing his sample size to 23, he produces statistical significance. Although Thompson in this example employs a large effect size, the same dynamic, i.e., different outcomes resulting from adding or loosing a few subjects can occur at any sample size. Carver (1978) confirms that "a mean difference that is small and not significant from a research standpoint can be <u>statistically</u> significant just because enough subjects were used in the experiment to make the result statistically rare under the null hypothesis" (p. 388).

The problem with blind reliance on significance testing is that it leads to misinterpretations of study findings. Suppose a researcher obtains a result that is statistically significant at the $p = .10$ or $p = .15$, but that with invariance testing would prove to be quite stable under sampling and thus generalizable to the population of interest. Suppose further that this researcher does not understand invariance testing and, consequently, lets a noteworthy result go unpublished. This unfortunate outcome would occur because an alpha level slightly less stringent than the commonly accepted $a = .05$ would be allowed to overshadow the importance of the generalizability of the findings. If science is the business of cumulating knowledge, then generalizability of study results warrents genuinely serious consideration.

The reverse is equally true. If this same researcher obtains a result significant at the $p = .01$ level, but again fails to conduct invariance testing and so fails to discover that this time the finding is sample specific and not generalizable, the same flawed thinking would be employed. Only this time the study would likely be written up and published. The fact that the finding does not apply to the population would likely go unnoticed, an unfortunate outcome of this scenario. Research conducted in this manner does not add to a body of knowledge and does not advance a field. To the contrary, a more likely result is that such practices retard development of a field because important, but statistically nonsignificant findings, are not included in the literature, while trivial, but statistically significant findings are.

There are critics of significance testing (Carver, 1978; Schneider & Darcy, 1984) who would agree with Thompson (1988, p. 100) that "significance is not...the end-all and be-all of research." Nonetheless, significance remains a paramount concern in research, causing Rosnow and Rosenthal (1989, p. 1277) to chide that "surely, God loves the .06 nearly as much as the .05. Can there be any doubt that God views the strength of evidence for or against the null as a fairly continuous function of the magnitude of p?" The admonition of Rosnow and Rosenthal (1989) notwithstanding, evidence indicates that such doubt does exist.

Carver (1978) notes that in 1977, despite rumblings against significance testing in the research community, only two of the 29 articles of empirical research published in the American Educational Research Journal did not use significance testing. This finding lead Carver (1978) to assert "apparently the case against such testing will have to be stated more loudly and more clearly to a wider audience if it is to have any effect" (p. 379). Twelve years later, in 1989, the present authors found that little had changed. Of 17 empirical articles published in the same journal, only one researcher found invariance testing important enough to warrant discussion, and only two others reported results that did not include levels of statistical significance.

Thompson (1987, 1988) reminds researchers that statistical significance testing does not provide information about the importance of results. For this reason, other analogs must be consulted for an indication of result noteworthiness. Carver (1978) concurs with this view, arguing that a major misconception of researchers involves the confusion of

statistical significance test results with findings regarding reproducibility.

Thoughtful researchers have begun to place importance on result replicability, which, in the view of Carver (1987), "is the cornerstone of science" (p. 392). This position on replicability is neither new nor novel. According to Tukey (1969), Sir Ronald Fisher, father of modern statistical testing, held the view that the "standard of firm knowledge was not one very extremely significant result, but rather the ability to repeatedly get results significant at 5%. Repetition is the basis for judging variability and significance and confidence. Repetition of results, each significant, is the basis, according to Fisher, of scientific truth" (p. 85).

Neale and Liebert (1986) corroborate the contention that replication is intrinsic to true scientific inquiry, stating "no one study, however shrewdly designed and carefully executed, can provide convincing support for a causal hypothesis or theoretical statement in the social sciences" (p. 290). Perhaps one reason for the difficulty in "exorcising the null hypothesis" is that researchers do not feel a suitable substitute has been offered. The present paper offers one such alternative--assessing the invariance of study results.

Invariance analysis provides more confidence that research results are stable and replicable across samples. The current study applied an invariance technique following a discriminant analysis. For readers unfamiliar with discriminant analysis, Huberty and Barton (1989) provide a very understandable explanation. Traditionally, four approaches have been used to assess the stability of discriminant function coefficients: (1)

the "empirical" method, (2) the "holdout" ("cross validation," or "split half") method, (3) the "Monte Carlo" method, and (4) the "random assignment" method. Daniel (1989) provides an explanation of each of these approaches. Other examples of invariance procedures following a discriminant analysis are provided by Jones (1989). The use of the Procrustean rotation invariance procedure is the focus of this paper. A concrete example is provided.

Heuristic Example

For the present paper, a discriminant analysis was calculated from a hypothetical data set with 64 cases and two predictor variables, X and Y. The first 32 cases were from a data set developed by Fish (1988). Four groups of 16 cases each were derived. The data set and the SPSSx commands for the discriminant analysis are presented in Tables 1 and 2, so that the reader is able to replicate and further explore the analysis.

INSERT TABLES 1 AND 2 ABOUT HERE

Procrustean rotation can be used with any multivariate technique. The name is derived from Greek mythology. Procrustes, a son of Poseidon, forced travelers spending the night at his home to fit his bed by either cutting off their legs or stretching their bodies. Similarly a Procrustean rotation forces orthogonal (uncorrelated) functions of factors to a "best fit" position after setting the factor vectors to unit length (1.0) "in order to equalize the contribution of each [factor vector] to the determination of the amount of rotation necessary" (Veldman, 1967, p. 238). This rotation technique can be used as a cross-validation

procedure, splitting the data from a single sample and comparing the factor vectors from each half.

The use of factor analysis as a means of validity evaluation is well known. Thompson and Pitts (1981/1982) describe this cosine application as a rotation of calculated factors to a position of "best fit" with a target matrix that has been theoretically derived. The target matrix determines how many factors are expected and the expected correlation between each item and each factor. Thus, the cosines of the angles between the actual and the hypothetical measures can be interpreted as validity coefficients.

In the present study, by splitting the original sample into two subsets, discriminant functions were generated resulting in two sets of coefficients for comparison using the "best fit" rotation method. Thompson (1986) provides a detailed review of this empirical method developed by Kaiser, Hunka, and Bianchini (1969) for "relating" factors derived from different samples of data. This method consists of projecting the two sets of factors into the same factor space and calculating the cosines of the angles among the factors across the two solutions. These cosines provide a measure of the relatedness of the two sets of factors, and are similar to correlation coefficients.

Thompson (1981, 1986) affirms that these coefficients in this application are analogous to test-retest coefficients and has called them invariance coefficients. They may also be utilized as adequacy coefficients for substantive interpretations. In this application function coefficients are submitted to a Procrustean rotation instead of structure coefficients, because the primary concern here is to investigate the similarity of the function equations used to produce function scores.

Table 3 presents these two sets of discriminant functions for the Table 1 data. Table 1 data contains the variable INVAR that was created to divide the data into eight groups of eight cases each. For this invariance procedure, odd groups were analyzed and their function coefficients used for Matrix A; Matrix B contained function coefficients from the analysis of the even groups.

INSERT TABLE 3 ABOUT HERE

Using the RELATE program by Veldman (1967), the standardized discriminant function coefficients from sample one are input as Matrix A and those from sample two as Matrix B. The decision regarding which matrix is to be designated the target for "best fit" rotation is arbitrary. Although the main focus of interest is the resulting matrix cosines, test r's should be first consulted. These evaluate the relation of the given variables from the two data sets within the factor space, and must be suitable for the functions being rotated to "best fit" to be suitable. The results of the rotation indicate a test r of .9993 and .9994 for each of the two variables, X and Y, respectively, indicating (since they are large) that the two discriminant functions can be rotated in this manner.

With respect to the resulting cosines among the functions, generally, to be considered replicated, functions should have a cosine of roughly .8 or higher (Thompson & Pitts, 1981/1982). Kaiser suggests .85 is reasonable and Gorsuch recommends results greater than .93 as exceptional; however, Thompson (1986) presents empirically derived cutoffs as an alternative to the theoretically derived cutoffs formulated by others.

The cosines among the functions across the solutions for these data are presented in Table 4. For this example, Matrix A, Function I has the "best fit" with Matrix B, Function II with a cosine of .8377; Matrix A, Function II has the best rotated fit with Matrix B, Function I with a cosine of .8377. These cosines are marginal and might be more meaningful intuitively were this heuristic example a real substantive study.

INSERT TABLE 4 ABOUT HERE

Relatively little is known about the characteristics of the various invariance estimates (Jones, 1989). Because of this, Thompson (1984) suggests that researchers employ several strategies in order to obtain both upper and lower bound estimates of the degree of capitalization on sampling specificity.

Summary

Thompson (1986) affirms that "researchers have increasingly recognized the critical nature of replication as the ultimate test of scientific findings and some have argued that replicability should replace significance testing as part of a new logic of truth testing" (p. 27). Even though the interpretation of invariance results remains a subjective judgment, this does not diminish the need for performing invariance procedures as an evaluation of replicability or generalizability of analytic results. Using two or more invariance procedures provides researchers an added measure of confidence regarding their results.

The present paper has elaborated one alternative applicable with all multivariate methods, i.e., Procrustean rotation. The RELATE computer

program in Veldman's (1967) book can be employed to implement the necessary calculations.

BIBLIOGRAPHY

Carver, R. P. (1978). The case against statistical significance testing. Harvard Educational Review, 48 (3), 378-399.

Cronbach, L. J. (1975) Beyond the two disciplines of scientific psychology. American Psychologist, 30, 116-127.

Daniel, L. G. (1989, January). Use of the jackknife statistic to establish the external validity of discriminant analysis results. Paper presented at the annual meeting of the Southwest Educational Research Association, Houston. (ERIC Document Reproduction Service No. ED 305 382)

Fish, L. (1986, November). The importance of invariance procedures as against tests of statistical significance. Paper presented at the annual meeting of the Mid-South Educational Research Association, Memphis.

Gorsuch, R. L. (1983). Factor Analysis (2nd ed.). Hillsdale, NJ: Lawrence Erlbaum Associates.

Huberty, C. J. & Barton, R. M. (1989). An introduction to discriminant analysis. Measurement and Evaluation in Counseling and Development, 22, 158-168.

Jones, G. (1989, January). Some examples of invariance procedures in discriminant analysis. Paper presented at the annual meeting of the Southwest Educational Research Association, Houston. (ERIC Document Reproduction Service No. ED 307 296)

Kaiser, H. F., Hunka, S., & Bianchini, J. (1969). Relating factors between studies based upon different individuals. In H. J. Eysenck and S. B. G. Eysenck (Eds.) Personality structure and measurement

(pp. 333-343). San Diego: Knapp.

Neale, J. M., & Liebert, R. M. (1986). Science and behavior: An introduction to the methods of research (3rd ed.). Englewood Cliffs, NJ: Prentice-Hall.

Rosnow, R. L. & Rosenthal, R. (1989). Statistical procedures and the justification of knowledge in psychological science. American Psychologist, 44 (10), 1276-1284.

Shulman, L. S. (1970). Reconstruction of educational research. Review of Educational Research, 40, 371-393.

Schneider, A. L. & Darcy, R. E. (1984). Policy implications of using significance tests in evaluation research. Evaluation Review, 8 (4), 573-581.

Thompson, B. (1981). Utility of invariance coefficients. Perceptual and Motor Skills, 52, 708-710.

Thompson, B. (1986, April). A partial test distribution for cosines among factors across samples. Paper presented at the annual meeting of the American Educational Research Association, San Francisco.

Thompson, B. (1987, April). The use (and misuse) of statistical significance testing. Paper presented at the annual meeting of the American Educational Research Association, Washington. (ERIC Document Reproduction Service No. ED 287 868)

Thompson, B. (1988). A note about significance testing. Measurement and Evaluation in Counseling and Development, 20 (4), 146-147.

Thompson, B. (1989). Statistical significance, result importance, and result generalizability: Three noteworthy but somewhat different issues. Measurement and Evaluation in Counseling and Development, 22, 2-6.

Thompson, B., & Pitts, M. C. (1981/1982). The use of factor adequacy coefficients. Journal of Experimental Education, 50, 101-104.

Tukey, J. W. (1969). Analyzing data: Sanctification or detective work? American Psychologist, 24, 83-91.

Veldman, D. J. (1967). Fortran Programming for the Behavioral Sciences. New York: Holt, Rinehart and Winston, 238-244.

Table 1
Hypothetical Data Set

Case	Group	X	Y	INVAR
1	1	4	2	5
2	1	5	3	8
3	1	4	4	2
4	1	4	5	3
5	1	3	4	4
6	1	6	5	6
7	1	5	6	7
8	1	7	5	2
9	1	6	6	1
10	1	8	6	8
11	1	7	6	1
12	1	9	7	5
13	1	8	7	4
14	1	8	8	3
15	1	9	8	7
16	1	9	9	6
17	2	1	2	8
18	2	3	3	4
19	2	3	5	3
20	2	3	5	6
21	2	2	5	5
22	2	4	6	4
23	2	4	5	2
24	2	5	6	5
25	2	6	6	6
26	2	6	6	1
27	2	6	7	7
28	2	7	7	8
29	2	7	7	2
30	2	8	9	3
31	2	8	9	7
32	2	9	9	1
33	3	4	1	8
34	3	4	2	6
35	3	3	2	3
36	3	2	4	5
37	3	5	3	2
38	3	7	4	1
39	3	4	5	7
40	3	5	4	5
41	3	7	5	8
42	3	9	5	6
43	3	6	5	4
44	3	5	6	1
45	3	7	6	7
46	3	9	7	3
47	3	8	6	5
48	3	8	5	2
49	4	1	7	4
50	4	1	2	3
51	4	1	1	2
52	4	2	2	8
53	4	2	3	3
54	4	2	3	1
55	4	3	2	7
56	4	3	3	4
57	4	3	4	7
58	4	4	5	6
59	4	4	4	5
60	4	4	5	4
61	4	4	6	2
62	4	5	6	1
63	4	5	7	8
64	4	5	7	6

Table 2

SPSSx Commands for Discriminant Analysis

```
FILE HANDLE MT/NAME='DISCRIMNT.DAT'
   DATA LIST FILE=MT/CASE 1-2 GROUP 7 X 12 Y 17
LIST VARIABLES CASE TO Y
DISCRIMINANT GROUPS=GROUP (1,4)
   /VARIABLES= X Y
   /STATISTICS=MEAN STDEV UNIVF RAW
```

Table 3

Matrices Entered Into Procrustean Rotation: Standardized Function Coefficients of Each Split-Sample Discriminant Analysis Run

MATRIX A (TARGET MATRIX, n = 32):

Function I	Function II
0.80530	1.53832
1.54426	0.79386

MATRIX B (n = 32):

Function I	Function II
1.56496	0.20726
1.34221	0.83097

Table 4

Cosines Among Factor Axes Resulting From Procrustean Rotation Invariance Procedure

A BY B	Function I	Function II
Function I	0.5462	0.8377
Function II	0.8377	0.5462